Increasing Object-Level Reconstruction Quality in Single-Image 3D Scene Reconstruction

Anna Ribic Antonio Oroz Meikel Kokowski Franz Srambical
Technical University of Munich
{firstname}.{lastname}@tum.de

Abstract

1. Introduction

While humans can easily infer the 3D structure as well as the complete (panoptic) semantics of a scene from a single image, this task has been a longstanding challenge in the field of computer vision. The task fundamentally prerequisites learning a strong prior of the 3D world. Traditional methods have made significant strides, from generating geometrically coherent structures [**?** **?**] to learning different instance semantics [**?** **?** **?**]. More recent approaches directly learn the 3D panoptic semantics as a whole [**?** **?**], yet they fall short in capturing the intricate details and nuances at the object level. This paper introduces a novel approach to bridge this gap by integrating a specialized object-level model into the reconstruction process, thereby leveraging the specialized model's object-priors.

2. Related Work

2D panoptic segmentation 2D panoptic segmentation merges semantic and instance segmentation, providing detailed pixel-level parsing of images, capturing both general categories (semantic segmentation) and individual object identities (instance segmentation) [**?**]. Since the original task formulation by **?**], a number of works have been proposed to solve the task [**?** **?** **?** **?** **?** **?** **?** **?** **?** **?** **?** **?** **?**], while more recent approaches [**?**] try to unify image segmentation in its entirety.

Single-view 3D reconstruction The work by **?**] was the first notable attempt at reconstructing 3D scenes from unordered photo collections. Since then, the field of image-based 3D reconstruction has seen a number of advancements, culminating in the task of single-view 3D reconstruction [**?** **?** **?** **?** **?** **?** **?**].

Shape priors **?**] note that the task of single-view 3D reconstruction is non-deterministic, as there are many 3D shapes that can explain a given single-view input, and propose to use shape priors to shape the solution space such that the reconstructed shapes are realistic, but not necessarily the ground truth.

3D scene understanding and panoptic reconstruction

Modality-conditioned shape generation 3D generative models represent objects in a variety of modalities, including point clouds [**?** **?**], occupancy grids [**?**], meshes [**?**], and signed distance functions [**?**]. Furthermore, these models can also be distinguished by the type of input they take, such as incomplete shapes [**?**], images [**?**], text [**?** **?**], or other modalities [**?**]. Notably, **?**] propose *SDFusion*, a 3D object reconstruction method conditioned on images, text and geometrical input.

DATASET

3. Method

4. Conclusion